Doble Grado en II+ADE 13/06/ 2018

Ecometría Examen Global Curso 2017-2018

DNI/NIE : APELLIDOS: Nombre:

Cada pregunta se resuelve en la hoja de su enunciado, no se pueden responder preguntas distintas en la misma hoja. Las respuestas se deben escribir con tinta azul o negra. Las respuestas deben ser breves pero razonadas. Errores conceptuales importantes pueden afectar a la calificación global del examen.

Teoría(3 puntos). 1) Obtener la expresión de la matriz de covarianzas del estimador $\hat{\boldsymbol{\beta}}$ del vector de coeficientes $\boldsymbol{\beta}$ del modelo de Gauss-Markov de la regresión lineal múltiple.

2) Explicar brevemente cual es la diferencia entre el test de significación t y el test de significación F en el modelo de Gauss-Markov de la regresión lineal múltiple, aclarando si en algún caso pueden ser equivalentes.

3) Definir serie temporal (débilmente) estacionaria. Poner un ejemplo de serie estacionaria y un ejemplo de serie no estacionaria.

Sol. 1) Está hecho en clase, lo recordamos:

Como

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (X'X)^{-1}X'\boldsymbol{\varepsilon},$$

y por las hipótesis del modelo de Gauss-Markov $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = var(\boldsymbol{\varepsilon}) = \sigma^2 I$, entonces

$$\text{var}(\hat{\boldsymbol{\beta}}) = E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})')] = (X'X)^{-1}X'E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']X(X'X)^{-1} = \sigma^2(X'X)^{-1}.$$

2) El test de significación F es una generalización del test de significación t cuando para obtener el modelo restringido se eliminan $g \geq 1$ variables explicativas

$$\text{H0} : \beta_{k-g+1} = \cdots = \beta_k = 0, \quad \text{H1} : \text{alguno de los parámetros } \beta_j \text{ es no nulo}, j = k-g-1, \ldots k.$$

Por tanto, el test t es equivalente al test F para $g = 1$: sólo se elimina una variable explicativa. La justificación estadística de que en ese caso el test F da el mismo resultado que el test t, es decir que los p-valores son los mismos (a pesar de que uno es de una cola y otro de dos), está en que

$$F_{1,n-k} = t^2_{n-k},$$

por la definición de las distribuciones correspondientes.

3) La definición de serie estacionaria está en la notas de clase. Ejemplos de series estacionarias son los ruidos blancos, por ejemplo,

$$\varepsilon_t = \text{NID}(0, 1).$$

Ejemplos de series no estacionarias son los caminos aleatorios, por ejemplo,

$$y_t = y_{t-1} + \text{NID}(0, 1).$$

Doble Grado en II+ADE **13/06/ 2018**

Ecometría Examen Global Curso 2017-2018

DNI/NIE : **APELLIDOS:** **Nombre:**

Ejercicio 1(2,5 puntos). Dados los siguientes datos

y	x_2	x_3
2	3	4
-2	2	6
8	5	2
6	4	3

1) Obtener el plano de regresión de y sobre x_2 y x_3. Dar un valor estimado de la covarianza de los estimadores de los dos coeficientes de x_2 y x_3.

2) Añadiendo y quitando variables, ¿cúal es el modelo significativamente más relevante?

Ayuda: $\begin{pmatrix} 4 & 14 & 15 \\ 14 & 54 & 46 \\ 15 & 46 & 65 \end{pmatrix}^{-1} = \begin{pmatrix} 697/3 & -110/3 & -83/3 \\ -110/3 & 35/6 & 13/3 \\ -83/3 & 13/3 & 10/3 \end{pmatrix}$

Sol. 1) A partir de los datos, en forma matricial

$$X = \begin{pmatrix} 1 & 3 & 4 \\ 1 & 2 & 6 \\ 1 & 5 & 2 \\ 1 & 4 & 3 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 2 \\ -2 \\ 8 \\ 6 \end{pmatrix},$$

primero observamos que la matriz X tiene rango $k = 3$, tomando por ejemplo el menor obtenido quitando la última fila, esto es necesario para tener un modelo de Gauss-Markov. Obtenemos

$$X'X = \begin{pmatrix} 4 & 14 & 15 \\ 14 & 54 & 46 \\ 15 & 46 & 65 \end{pmatrix}, \quad \hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y} = \begin{pmatrix} 8/3 \\ 5/3 \\ -4/3 \end{pmatrix}.$$

Plano de regresión: $y = \frac{8}{3} + \frac{5}{3}x_2 - \frac{4}{3}x_3$.

Una estimación de la matriz de covarianzas $\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(X'X)^{-1}$ viene dada por $s^2(X'X)^{-1})$. Luego una estimación de $\text{cov}(\hat{\beta}_2, \hat{\beta}_3)$ es $s^2 a_{23}$. Basta calcular s^2,

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - X\hat{\boldsymbol{\beta}} = \begin{pmatrix} -1/3 \\ 0 \\ -1/3 \\ -2/3 \end{pmatrix}, \quad s^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-k} = \frac{2}{3} \Rightarrow s^2 a_{23} = \frac{26}{9},$$

que es la estimación pedida.

2) Tenemos cuatro modelos posibles: modelo constante, modelo completo, modelo restringido con variables y, x_2 y modelo restringido con variables y, x_3. Hay un test F global que compara el model completo y el constante y dos tests t que comparan el completo con los dos restringidos, respectivamente.

TEST F GLOBAL:

H0: $\beta_2 = \beta_3 = 0$, H1: alguno de los parámetros β_2 o β_3 es no nulo.

Calculemos el F-valor:

$$SST = 59, \quad \hat{\varepsilon}'\hat{\varepsilon} = \frac{2}{3}, \quad R^2 = 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{SST} = 175/177, \quad F = \frac{1}{2}\frac{R^2}{1-R^2} = \frac{175}{4} = 43.75.$$

Este valor cae en la región de no rechazo $(0, 199.5)$ del estadístico $F_{2,1}$; por tanto, no rechazamos la hipótesis H0, y el modelo constante es más significativo que el completo: podemos prescindir de las dos variables explicativas conjuntamente.

TESTS t: esperamos que los tests t corroboren lo anterior.

Aquí tenemos dos tests de significación independientes, uno para cada variable, comparando el modelo completo con el restringido al quitar una de las variables:

a) Test variable x_2: H0: $\beta_2 = 0$, H1: $\beta_2 \neq 0$.

b) Test variable x_3: H0: $\beta_3 = 0$, H1: $\beta_3 \neq 0$.

Calculando los t-valores:

$$s_2 = \frac{\sqrt{35}}{3}, \; s_3 = \frac{\sqrt{20}}{3}, \; t_2 = \frac{\hat{\beta}_2}{s_2} = 0.8451, \; t_3 = \frac{\hat{\beta}_3}{s_3} = -0.8944.$$

Ambos valores caen en la región $(-12.706, 12.706)$ de no rechazo del estadístico t_1. Por tanto, *el modelo constante es claramente el más significativo.*

Doble Grado en II+ADE 13/06/ 2018

Ecometría Examen Global Curso 2017-2018

DNI/NIE : APELLIDOS: Nombre:

Ejercicio 2 (1.5 puntos)

1) Supongamos que la ecuación en diferencias homogénea de orden dos

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2},\ k \in \mathbb{N}$$

es tal que su ecuación característica tiene raíces complejas, $\lambda = |\lambda| e^{i\omega}$, $\overline{\lambda} = |\lambda| e^{-i\omega}$ (sabemos que han de ser complejos conjugados). Demostrar que, entonces, la solución general de la ecuación anterior en diferencias se puede excribir como

$$\rho_k = A|\lambda|^k \cos \omega k + B|\lambda|^k \sin \omega k,$$

siendo A y B constantes *reales*.

2) Dado el proceso aleatorio

$$y_t = 2y_{t-1} - 2y_{t-2} + \mathrm{NID}(0, 0.2), \quad t \in \mathbb{Z}.$$

Obtener sus esperanzas, varianzas y sus funciones de autocorrelación.

Sol.

1) Sabemos, por la teoría de clase, que podemos escribir la solución como

$$\rho_k = C|\lambda|^k e^{ik\omega} + \overline{C}|\lambda|^k e^{-ik\omega},$$

es decir, si $C = re^{i\varphi}$, utilizando la fórmula de Euler, y teniendo en cuenta que el coseno es una función par y el seno impar, obtenemos

$$\rho_k = re^{i\varphi}|\lambda|^k e^{i\omega k} + re^{-i\varphi}|\lambda|^k e^{-i\omega k} = r|\lambda|^k (e^{i(\varphi+\omega k)} + e^{i(-\varphi-\omega k)}) =$$
$$r|\lambda|^k [\cos(\varphi + \omega k) + i \sin((\varphi + \omega k) + \cos(\varphi + \omega k) - i \sin((\varphi + \omega k)] =$$
$$2r|\lambda|^k \cos(\varphi + \omega k) = 2r|\lambda|^k (\cos\varphi \cos \omega k - \sin\varphi \sin \omega k) = A|\lambda|^k \cos \omega k + B|\lambda|^k \sin \omega k,$$

para constantes reales $A = 2r \cos\varphi$ y $B = -2r \sin\varphi$.

2) La ecuación caracteristica es

$$\lambda^2 - 2\lambda + 2 = 0,$$

con autovalores

$$\lambda_1 = 1 + i,\ \lambda_2 = 1 - i.$$

Como estos autovalores complejos son tales que sus módulos están fuera del círculo unidad, $|\lambda_i| > 1$, entonces la serie no es estacionaria, y no se pueden calcular sus esperanzas, varianzas y sus funciones de autocorrelación con las fórmulas dadas en clase (Observación: se ha considerado como correcto este apartado si se han sabido obtener los autovalores).

Doble Grado en II+ADE 13/06/ 2018

Ecometría Examen Global Curso 2017-2018

DNI/NIE : APELLIDOS: Nombre:

Ejercicio 3 (con ordenador) (3 puntos) Descargar los ficheros BudgedFood.rda y BudgedFood.pdf.

Vamos a estudiar cómo depende de otras variables, el porcentaje del consumo total del hogar que dedicaban los hogares españoles a la alimentación en 1980. El fichero de datos es BudgetFood.rda (es un fichero con extensión .rdata, con lo que *se puede cargar directamente en el RCommander: Datos-> Cargar conjunto de datos...*). La información de qué significan los datos está en el fichero BudgedFood.pdf. De todas las variables que aparecen en el fichero descartaremos el género ("sex"). Por tanto, queremos estudiar la dependencia de la variable "wfood" respecto al resto de variables del fichero de datos, *menos el género*. Este es el modelo que consideraremos como completo.

1) Escribir e interpretar los resultados de la regresión del modelo completo: plano de regresión, estimación de los parámetros, coeficiente de determinación, errores estándar, t-valores, p-valores, etc.

2) ¿Es significativa la siguiente afirmación?:

"El número de miembros del hogar influye - céteris páribus - en el porcentaje del consumo que dedican a la alimentación, diez veces más que la edad del (o de la) cabeza de familia"

Obtener el resultado de dos formas:

a) mediante el p-valor que obtiene el programa,

b) mediante el F-valor que obtiene el programa (y las tablas).

Sol. 1) Denotando las variables explicativas totexp, age, size y town como x_{tot}, x_{age}, x_{siz} y x_{tow}, respectivamente. Ajustando el modelo mediante regresión lineal, descartando el intercepto, obtenemos

	$\hat{\beta}_j$	s_j	t_j	p-valor
x_{age}	$2.135 \cdot 10^{-3}$	$6.228 \cdot 10^{-5}$	34.28	$< 2 \cdot 10^{-16}$
x_{siz}	$2.199 \cdot 10^{-2}$	$5.446 \cdot 10^{-4}$	40.38	$< 2 \cdot 10^{-16}$
x_{tot}	$-1.378 \cdot 10^{-7}$	$1.524 \cdot 10^{-9}$	-90.43	$< 2 \cdot 10^{-16}$
x_{tow}	$-1.829 \cdot 10^{-2}$	$7.336 \cdot 10^{-4}$	-24.94	$< 2 \cdot 10^{-16}$

$s = 0.1338$, $R^2 = 0.348$, $\bar{R}^2 = 0.3478$,

test F de significación global con $F(4, 23967)$: F-valor= 3197, p-valor$< 2.2 \cdot 10^{-16}$.

Hay un test de significación global:

H0:$\beta_{age} = \beta_{siz} = \beta_{tot} = \beta_{tow}$, H1: alguno de los β anteriores es no nulo

y cuatro tests de t de significación individuales:

H0: $\beta_j = 0$, H1: $\beta_j \neq 0$, $j = age, ..., tow$.

Lo primero que llama la atención es que todos los p-valores tanto los de los tests t, como el

F global son, a efectos prácticos, nulos. Por tanto, el modelo adecuado es claramente el modelo completo. El hecho de que R^2 que no esté cerca de la unidad, aunque los p-valores son muy pequeños, se explica por el elevado número de observaciones.

También resaltamos que hay dos variables: el presupuesto total del gasto y el tamaño de la ciudad que influyen negativamente, céteris páribus, en el porcentaje que se dedica a la alimentación del hogar, aunque en el caso del presupuesto muy débilmente (pero no se puede considerar que no influye nada: ¡el p- valor es prácticamente nulo!). Una posible interpretación de la influencia negativa de los hogares del mundo rural es que en los hogares del mundo rural una parte de su alimentación provenía de recursos del propio hogar, aunque se podrían encontrar otras. La influencia negativa del presupuesto total parece clara, pues a mayores ingresos menor porcentaje del mismo dedicado a la alimentación, céteris-páribus: para hogares del mismo tamaño, etc.

2) El contraste es mediante un test F:

H0: $\beta_{siz} - 10\beta_{age} = 0$, H1: $\beta_{siz} - 10\beta_{age} \neq 0$

Construyendo un nuevo modelo, a partir del completo, teniendo en cuenta que:

$$\beta_{siz} = 10\beta_{age} \Rightarrow \beta_{age}x_{age} + \beta_{siz}x_{siz} + \beta_{tot}x_{tot} + \beta_{tow}x_{tow} = \beta_{age}z + \beta_{tot}x_{tot} + \beta_{tow}x_{tow},$$

siendo $z := x_{age} + 10x_{siz}$.

Hacemos el ajuste de este nuevo modelo con las variables explicativas z, x_{tot} y x_{tow}, y contrastamos este modelo con el modelo completo: Modelos->Test de hipótesis->Comparar dos modelos. Del resultado nos interesa el F-valor y el p-valor, que son 0.8392 y 0.3596, respectivamente. Como el p valor es claramente mayor que 0.05, no se puede rechazar la hipótesis H0. Utilizando el F-valor, como para F(1,23968), $23968 \simeq \infty$ a efectos prácticos y 0.8392 cae claramente en la zona de no rechazo (0,3.842), obtenemos la misma conclusión: la afirmación

"El número de miembros del hogar influye - céteris páribus - en el porcentaje del consumo que dedican a la alimentación, diez veces más que la edad del (o de la) cabeza de familia",

es, sin lugar a dudas, significativa.